L'entrelacement lexical des textes. Cooccurrences et lexicométrie

Damon Mayaffre

L'ENTRELACEMENT LEXICAL DES TEXTES, CO-OCCURRENCES ET LEXICOMÉTRIE[*1]

Damon Mayaffre
CNRS - UMR Bases, Corpus et Langage - MSH de Nice-Sophia-Antipolis

1 INTRODUCTION

La linguistique de corpus, célébrée chaque année à Lorient, entend adjoindre à l'introspection linguistique l'observation de faits langagiers réels consignés en corpus. Pour cette raison, le texte (seul objet linguistique empirique puisque nous ne nous exprimons ni en mots désincarnés, ni en phrases indépendantes) est son objet favori. Pour cette raison encore, les méthodes de traitement deviennent centrales, puisqu'une fois constituées en corpus (corpus numériques de plus en plus gros) les données ne sauraient être traitées par la seule intuition. Pour cette raison enfin, dans le domaine sémantique, la *co(n)textualisation* des formes nous semble la philosophie même de nos pratiques, étant entendu que le sens des choses n'émerge qu'en co(n)texte, qu'en corpus.

Cette contribution entend donc réfléchir aux méthodes de co(n)textualisation proposée par la lexicométrie pour traiter les grands corpus textuels.

La volonté de remettre les formes dans leur contexte[2] se caractérise, en lexicométrie, par deux types de comportement de l'analyste et deux modes de fonctionnalités classiques des logiciels : le *retour au texte*, simple mais essentiel, et le développement d'une *statistique contextualisante*, syntagmatique ou co-occurrencielle.

1.1 Le retour au texte

Dans les logiciels reconnus sur le marché scientifique tels *Hyperbase*, *Lexico*, *Astartex* ou *Weblex*, le retour direct au texte plein ou le retour indirect, partiel mais organisé au texte par l'intermédiaire de concordances est systématique à toutes les étapes du traitement. Ainsi pourra-t-on par simple clic accéder au texte intégral *via* une liste de spécificités ou une constellation de mots disposés sur une Analyse Factorielle des Correspondances (AFC), et, par là, naviguer dans le corpus pour le lire de façon naturelle. De la même manière, on le sait, il sera possible de convoquer en un instant toutes les phrases ou tous les paragraphes contenant telle ou telle forme puis organiser ces *concordances* et en permettre une lecture aisée. Les occurrences –seules et désincarnées– sont donc les entrées utiles et nécessaires pour le traitement lexicométrique mais le retour au (co)(n)texte est posé comme la condition de l'interprétation. Grâce à l'hypertextualité, les logiciels organisent des parcours de lecture :

---

[*] Pour faire référence à cet article : Mayaffre Damon, « L'entrelacement lexical des textes, co-occurrences et lexicométrie », revue électronique *Texte et corpus*, n°3 / août 2008, Actes des Journées de la linguistique de Corpus 2007, p. 91-102
(disponible sur http://web.univ-ubs.fr/corpus/jlc5/ACTES/ACTES_JLC07_mayaffre.pdf)

[1] Cette contribution sera reprise et augmentée dans *Sémantique et Syntaxe*, n°9, 2008 (à paraître).

[2] Ont été notés ci-dessus « co(n)textualisation » et « co(n)texte ». Renonçons désormais, sauf exception, à alourdir le texte des parenthèses, étant entendu qu'on traitera toujours du *contexte linguistique* aussi appelé *co-texte*. Plus précisément, c'est le *co-texte immédiat* qui sera le plus souvent considéré, bien que la contextualisation linguistique d'une forme ne saurait s'arrêter à la phrase ou au paragraphe pour s'élargir au texte et au corpus (*Infra*).

ceux-ci sont certes originaux au sens où ils font appel à des ressources informatiques et à une navigation inaccessible aux pratiques manuelles-oculaires, mais l'acte de lecture est ici lui-même traditionnel au sens d'une confrontation entre le texte dans sa chaîne contextuelle naturelle et l'analyste. La lexicométrie est le bras armé de la linguistique de corpus et de l'*herméneutique numérique* qui voit aujourd'hui le jour (par exemple Viprey, 2005a et b ; Mayaffre, 2002a et 2006) : organiser le retour au texte pour en permettre la lecture et favoriser l'acte final interprétatif est une de ses tâches fondamentales.

1.2 La statistique contextualisante

Mais retourner trop vite à une lecture traditionnelle du texte, c'est renoncer trop tôt aux traitements quantitatifs dont la lexicométrie postule la pertinence pour lire, comprendre, interpréter les grands corpus textuels. Pour cette raison, le traitement statistique occurrentiel d'essence lexicographique – les occurrences nucléaires, décontextualisées qui une fois comptées, triées, indexées sont censées *faire référence* et renvoyer à des ontologies – se prolonge par des traitements statistiques ou mathématiques contextualisants de type co-occurrenciels et d'essence lexicologique. En amont des travaux en cours de Mellet & Barthélemy(2007) ou Luong, Longrée & Mellet (2008) sur la *topologie textuelle* dont les prétentions de modélisation de la textualité sont plus importantes, c'est de ce traitement co-occurrentiel que cette contribution veut traiter, après les travaux pionniers, en France, de Demonet *et al.* (1975), Tournier (1980), Lafon (1984) et ceux plus récents de Viprey (1997, 2005a et b), Heiden (1998 et 2004), Véronis (2003 et 2004), Martinez (2003) ou Brunet (2006, 2007 et 2008)[3].

2 COLLOCATION, CORRÉLATION, CO-OCCURRENCE

Le terme « co-occurrence » serait bien établi en Analyse de Données Textuelles (ADT), si certains auteurs, venant d'autres horizons, ne s'appliquaient à en brouiller le sens : la co-occurrence est la co-présence ou *présence simultanée* de deux unités linguistiques (deux mots par exemple ou deux codes grammaticaux) au sein d'un même contexte linguistique (le paragraphe ou la phrase par exemple, ou encore une fenêtre arbitraire). Cette co-occurrence peut être grossièrement constatée, puis vainement exprimée, en fréquence absolue. Mais plus pertinemment, la lexicométrie la constate et l'exprime grâce à des coefficients statistiques à même de mesurer le degré de significativité des co-présences ou attractions trouvées. Nombre de modèles et de coefficients ont été à ce jour proposés : Lafon (1984), Church & Hanks (1990), Dunning (1993), Fung & McKeown (1997), Manning & Schütze (1999), Véronis (2003 et 2004), Wu & Zhou (2003), etc. Et nous rappelons en note à la suite de Brunet (2007 et 2008) le mode de calcul hypergéométrique d'influence saint-clousienne implémenté dans *Hyperbase* qui sera utilisé ici[4].

Au moins deux termes complémentaires ou concurrents (la *collocation* et la *corrélation*) permettent de préciser les choses pour souligner la dimension générique de la co-occurrence.

[3] Les contributions sur la co-occurrence se heurtent au problème bibliographique : depuis Firth (1957) et Harris (1957), les articles traitant directement ou indirectement de la co-occurrence sont trop nombreux pour être synthétisés. Nous nous excusons de la multiplication des références qui accompagneront le propos.

[4] Soit s = nombre de phrases ou de paragraphes, f = fréquence du mot-pôle dans le texte, g = fréquence du mot co-occurrent dans le texte et k =co-occurrence observée. Alors :
Prob (x=k) = (f ! (s+g) ! g ! (f+s) !) / (k ! (f-k) ! (g-k) ! (s+k) ! (f+g+s) !)

2.1 Co-occurrence et collocation

Le terme « collocation » apparaît parfois comme synonyme de co-occurrence, particulièrement dans la littérature anglo-saxonne (Williams, 1999 ; Daille & Williams, 2001). Les collocations pointent pourtant le plus souvent des co-occurrents d'un certain type, ceux qui entretiennent des relations syntaxiques (ou parfois distributionnelles). C'est en ce sens que Haussmann (1979) ou Mel'Cuk *et al.* (1984), pour ne citer que les exemples les plus célèbres, les utilisent. C'est ainsi que les définissent Béjoint & Thoiron (1992, p. 517) :

> Les collocations sont des associations privilégiées de quelques mots (ou termes) reliés par une structure syntaxique et dont les affinités syntagmatiques se concrétisent par une certaine récurrence en discours.

Par là, si la co-occurrence prétend constater statistiquement les usages individuels – des associations libres relevant du choix du locuteur –, la collocation nous renvoie linguistiquement déjà du côté des *contraintes* du système (ou plutôt *des* systèmes idiomatiques) : indéniablement, certaines co-occurrences apparaissent comme des faits de langue, évidemment dans des lexies composées (*chemin de fer*) qui sont hors du champ mais aussi dans les syntagmes semi-figés (*semi-fixed combinations*), objets favoris des études collocatives (*pluie battante*, *salaire de misère*, *gravement malade*, etc.). De fait, la recherche des collocations – qui apparaît donc comme une sous-espèce spécialisée de celle des co-occurrences – est le plus souvent tendue vers la mise à jour, à finalité linguistique, des expressions idiomatiques, des unités phraséologiques, des phrasèmes ou semi-phrasèmes, des locutions, etc. Nous retrouvons ainsi la notion de collocation particulièrement présente dans les travaux de traductologie (automatique) afin de déterminer des formules propres à une langue qui exigent un effort de traduction particulier loin du mot à mot. Dans ce cadre, la recherche de collocations aboutit en général aux antipodes de notre propos c'est-à-dire à des travaux lexicographiques (*versus* lexicologiques), avec l'établissement de nomenclatures ou de dictionnaires censés consigner des locutions, leur sens définitif et éventuellement leur traduction dans d'autres langues : ainsi pourra-t-on consulter pour l'anglais, l'allemand ou le français, les dictionnaires de Benson *et al*. (1997), Ilgenfritz *et al*. (1989), Mel'Cuk *et al*. (1984) ou, plus strictement pour les mots composés, le DELAC de Silberztein et Gross (1996).

2.2 Co-occurrence et corrélation

Le terme « corrélation » pourrait présenter l'avantage de suggérer l'approche statistique (*cf.* les *indices de corrélation* en vigueur dans tout modèle statistique). Là où la co-occurrence de deux termes (si elle est exprimée naïvement en valeur absolue) peut être marginale et fortuite, leur corrélation semble souligner l'intensité de la relation ; la corrélation serait ainsi une co-occurrence *significative* d'un point de vue statistique.

Pourtant, dans la littérature, la corrélation pointe une autre réalité. Comme la collocation, la corrélation stigmatise des co-occurrents d'un certain type, ceux qui entretiennent une relation sémantique. Deux corrélats seraient deux co-occurrents qui ont une relation de sens.
Outre le fait que, de manière intriquée – donc problématique, un corrélat (sémantique) peut être un collocat (syntaxique) et vice versa, la notion de corrélation (comme celle de collocation) présente un danger de confusion épistémologique entre les différents plans de l'analyse : nous passons en effet imperceptiblement du *constat* de co-présence statistique (la co-occurrence) à la *signification* linguistique de cette relation (le corrélat). Précisons bien : notre propos immédiat consistera à montrer que la co-occurrence porte en elle un potentiel important pour la sémantique de corpus, la science et l'interprétation du texte : c'est en passant de l'occurrence à la co-occurrence que la lexicométrie accède à la lexicologie et que l'ADT entre dans la sémantique interprétative. Mais nous tenons à distinguer ce qui relève de la description formelle ou matérielle d'un phénomène, et le sens toujours négociable à donner

à celui-ci : parler de *corrélats sémantiques* à propos de *co-occurrences statistiques*, c'est conclure ce que nous voulons ici postuler.

3 LA CO-OCCURRENCE : SA DIMENSION HERMÉNEUTIQUE ET SES ENJEUX POUR L'ADT

La lexicologie est l'étude des vocabulaires en usage Eluerd (2000) : la contextualisation des vocables en est la clef. La linguistique est l'étude des textes définis comme le seul objet empirique du linguiste (Adam, 1999 ; Rastier, 2001) : la contextualisation des unités ou grandeurs textuelles, au sein de parcours de lecture contrôlés, en est ici encore la clef.

Si elle veut servir le vocabulaire et le texte, la lexicométrie doit donc proposer des outils pour traiter du contexte ; le contexte étant défini autant comme un environnement matériel bien circonscrit (une fenêtre) que comme un moment où le sens prend forme ou un lieu virtuel, éventuellement discontinu, où la textualité prend corps.

3.1 Au palier supérieur

Au palier supérieur, le contexte est non seulement *tout le texte* – selon l'expression de François Rastier –, mais encore le corpus textuel dans son ensemble, macro objet qui informe ses composants : c'est en effet, pour finir, au sein du corpus que les mots, les phrases, les textes prennent sens pour l'analyste ; c'est au sein du corpus que s'explicitent et s'organisent des stratégies de lecture interprétatives[5].

À vrai dire, cette affirmation – la place centrale du corpus – qui paraît faire aujourd'hui l'unanimité en linguistique (voir récemment, même pour la phonologie, le point de vue de Laks (2008)) est LE postulat originel des pratiques lexicométriques. Fondamentalement, la lexicométrie, au service de la linguistique de corpus, s'est présentée comme une alternative complémentaire à la linguistique introspective en donnant aux linguistes les moyens qui leur manquaient de mettre à l'épreuve leur intuition, et d'observer de manière critique les grands corpus empiriques afin de repérer les régularités linguistiques, les caractéristiques ou anomalies du langage réel : il y a là une posture essentielle sur laquelle nous ne pouvons revenir. Techniquement surtout, c'est bien le corpus dans son ensemble qui constitue la norme statistique sans laquelle aucun décompte ne ferait sens. Le dispositif de traitement tout entier repose, on le sait, sur l'idée de *norme quantitative endogène* au corpus[6]. La fréquence locale d'un mot dans une partie du corpus est mise en rapport avec la fréquence totale dans le corpus. C'est seulement par la mise en contraste des parties du corpus, et sur le postulat que l'ensemble du corpus représente une norme ou un étalon cohérent, que repose le traitement quantitatif. Si certains linguistes hésitent encore à admettre que le sens est différentiel, la statistique, elle, ne peut fonctionner sans cette affirmation : on ne jugera la fréquence du mot

[5] Nous ne reprendrons pas ici la réflexion sur le rôle et la place du corpus en linguistique. Nous renvoyons le lecteur à la revue dédiée à la question : *Corpus*. Résumons pour les sciences du texte : les corpus sont des objets construits – donc critiques et problématisés – qui informent leurs composants. Nous avons souligné que cette information était d'autant plus performante que les corpus étaient *réflexifs* (Mayaffre 2002b et 2006), c'est-à-dire susceptibles d'internaliser leurs ressources interprétatives : en miroir, les textes du corpus doivent s'éclairer mutuellement ; se *réfléchir* les uns les autres ; chacun d'entre eux constituant le co-texte immédiat de tous, et l'ensemble, l'intertexte de chacun.

[6] Impossible ici aussi de revenir sur ce postulat fondamental : il n'existe pas de fréquence en langue mais seulement en corpus. La nécessité statistique de travailler sur un corpus clos et étalon s'appuie, précisément, sur une conscience lexicologique. Le vocabulaire ne peut être abordé qu'en usage, c'est-à-dire *en corpus* : notre objet est le vocabulaire du corpus, non le lexique du dictionnaire. *Endogène* est sans doute le mot-clef de la linguistique de corpus, ainsi peut-on envisager une statistique endogène, une stylistique endogène (Viprey 1997), une lexicologie endogène, une sémantique endogène, etc.

« autorité » dans le discours de Sarkozy durant la campagne électorale 2007 comme importante, signifiante, porteuse de sens, qu'au regard de la norme que propose par exemple le corpus des discours de tous les candidats à l'élection présidentielle.

3.2 Au palier inférieur

Le contexte, linguistique comme statistique, est donc au palier supérieur le corpus. Au palier inférieur, nous voulons poser que *le contexte minimal d'un terme est la co-occurrence*. Nous considérerons en effet, en corpus, que la forme minimale du contexte d'un terme, nécessaire à sa compréhension-interprétation, n'est pas le syntagme ou la phrase mais la co-occurrence ; ou, dit autrement encore, nous définirons ici la co-occurrence comme la forme minimale du contexte qui présente l'avantage de se trouver accessible de manière systématique, étant entendu que nous ne saurions considérer, même avec un concordancier, un par un, tous les mots dans toutes leurs chaînes. Pour Saussure (1995, p. 150 et ss.), chaque occurrence est un hapax dont la valeur diffère avec le contexte, l'intonation, etc. : comment faire alors pour réaliser une synthèse des usages ou même seulement relever la moindre régularité ? Pour Guiraud (1960, p. 19), le sens d'un mot « se définit finalement par la somme de ses emplois » : mais comment faire pour *sommer* des emplois linguistiques (surtout lorsqu'il s'agit d'hapax) ? À mi-chemin, le traitement des co-occurrences entend considérer tous les mots en leurs (con)textes, et extraire de manière systématique de ceux-ci les formes significativement associées à ceux-là ; ou encore, considérer tous les paragraphes du corpus et y repérer systématiquement les associations linguistiques récurrentes jugées comme significatives et, pour cette raison, actrices principales de la textualité.

Concrètement en tout cas, constater que *a* et *b* sont co-occurrents n'est rien d'autre, pour nous, que contextualiser minimalement l'un par l'autre ; déterminer l'ensemble *b, c, d, e,* etc. des co-occurrents de *a,* c'est définir l'ensemble des contextes minimaux (mais pertinents) de *a* au sein du corpus.

Enfin, signalons que la définition de la co-occurrence comme forme particulière du contexte aboutit à un glissement de la vision habituelle du contexte, ne serait-ce que parce qu'il ne s'agit plus ici de chaîne, de fenêtre, de suite continue mais d'*associations* ; associations discontinues le plus souvent, parfois transphrastiques. Ici la définition matérielle du contexte – le contexte c'est avant tout un co-texte entendu comme un environnement textuel immédiat et contigu – se trouve équilibrée par une dimension herméneutique – le contexte, *c'est ce qui fait sens* ; ce qui sémantise un terme en autorisant l'interprétation. La co-occurrence ainsi définie serait alors une forme indiquée pour incarner le *passage* –passage minimal évidemment– dans la double dimension que lui donne Rastier (2007, p. 30) (« *extrait* de l'expression » et « *fragment* du contenu »)[7].

4 DES CO-OCCURRENCES DE « PATRIE » CHEZ SARKOZY

Essayons d'illustrer la réflexion théorique par une étude de cas. L'analyse porte sur le discours de la campagne électorale française de 2007. Deux corpus seront traités, le premier est composé des principaux discours des principaux candidats du premier tour à l'élection présidentielle : Laguiller, Buffet, Royal, Bayrou, Sarkozy et Le Pen. Le second contient l'exhaustivité des discours de meeting – 34 exactement – de Sarkozy du 1er janvier 2007 à la veille de son élection ; l'ensemble de ces discours, représentant plus d'un million de mots, est

[7] Pour un développement de cette idée, *cf.* la version augmentée de notre propos dans *Sémantique et Syntaxe*, n°9, 2008.

disponible sur le site *Discours 2007* de Jean Véronis (http://sites.univ-provence.fr/veronis/Discours2007/).

L'objectif est donc de stabiliser, *via* la recherche des co-occurrences, des contextes minimaux d'utilisation d'un terme afin de déceler des isotopies sémantiques et nourrir l'interprétation.

4.1 Constat occurrentiel

La démonstration part d'un premier constat lexicométrique, d'ordre seulement occurrenciel : la distribution du mot « patrie » dans le corpus du premier tour.

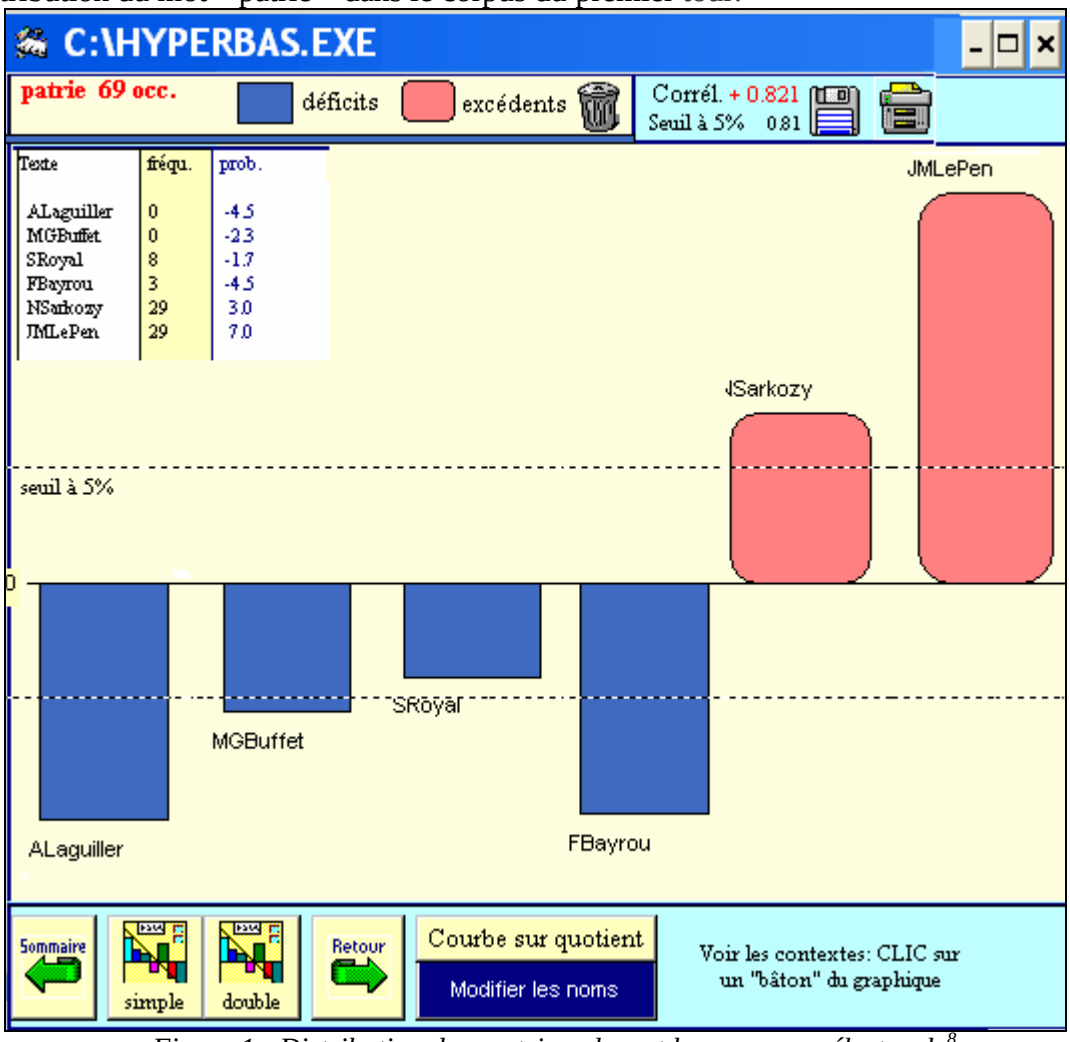

Figure 1 : Distribution de « patrie » durant la campagne électorale[8]

Ce constat est suggestif : plus on est à droite de l'échiquier politique plus on emploie « patrie », plus on est à gauche moins on l'utilise. Dans cette grille de lecture politique du corpus, seul Bayrou fait exception. Néanmoins, on ne saurait ici aller plus loin dans l'interprétation textuelle et l'inférence socio-linguistique, sauf à supposer trop vite un sens

[8] Les sous-utilisations et sur-utilisations mesurées par rapport à la norme du corpus sont ici exprimées en écart réduit. Les non-spécialistes de lexicométrie ne s'étonneront pas, par exemple, que la sous-utilisation de « patrie » soit plus marquée chez Laguiller (-4,5) que chez Buffet (-2,3) alors même que l'une et l'autre n'utilisent pas le terme (fréquence = 0). Cela s'explique par le fait que le sous-corpus de Laguiller est plus important que celui de Buffet : l'absence de « patrie » en devient plus significative.

déjà-là à « patrie », là où l'on sait que le mot possède un des signifiés les plus problématiques de la langue politique française.

4.2 Le sens de « patrie » : rappel historique

Il en va en effet ainsi de quelques termes dans le discours politique. « Patrie », comme « peuple » par exemple, est dans la bouche d'un homme politique contemporain, à l'image des éléments des rêves, sémantiquement *surdéterminé*. Sans doute peut-on parler de polysémie linguistico-politique en dépit du *Petit Robert* qui accorde une seule entrée et un seul sens au mot.

Tout au contraire, dans le bien nommé *Dictionnaire des usages socio-politiques*, Guilhaumou & Monnier (2006) montrent que dès l'origine moderne révolutionnaire, le mot, en discours, se charge d'acceptions différentes. On en distingue habituellement trois :

– une acception territoriale : la patrie c'est le territoire, la frontière à défendre contre les armées étrangères à partir de 1792, la terre des pères (voire la race).

– une acception politique : la patrie, en France, c'est la république des sans-culottes *versus* la monarchie ; par là ce sont des valeurs politiques comme la démocratie, la liberté, l'égalité, la vertu. Loin de la terre, la patrie c'est donc l'Idée (l'idée républicaine s'entend).

– une acception sociale enfin, en lien avec l'acception précédente : la patrie ce sont les crève-la-faim, la paysannerie pauvre et le peuple contre l'aristocratie, les riches, les privilégiés de la noblesse. La patrie des révolutionnaires ce n'est pas seulement l'idée républicaine naissante mais une revendication de justice sociale.

Dans le « allons enfants de la patrie » de La Marseillaise (25 avril 1792), dans *l'événement discursif* (Guilhaumou, 2006) majeur que constitue la déclaration par la Législative de « la patrie en danger » (11 juillet 1792), dans le « vive la patrie !» de l'armée révolutionnaire de Valmy partant au combat (20 septembre 1792), ces trois dimensions sont présentes. À l'inverse, l'épisode de Coblence représente, sur ces trois points, le symbole de l'anti-patrie.
Cette polysémie, originelle donc, ne fait ensuite que se complexifier au fil du temps : tout au long du XIX^{e} et du XX^{e} siècles, le mot prend alors une épaisseur historico-sémantique insondable.

Bref, sur ce substrat, il est facile de comprendre que l'occurrence seule de « patrie » dans le discours de Sarkozy ne signifie rien ; et le recours au *Robert* ou au *Larousse* risquerait de nous égarer définitivement. Seule, ici comme ailleurs, l'approche contextualisante, *en corpus*, peut instruire le débat, d'autant que la plupart des sens historiques qu'un dictionnaire peut consigner ont muté dans la France du XXI^{e} siècle. La lexicologie est une pratique endogène à un corpus ou elle n'est pas.

4.3 Co-occurrences de « patrie » et isotopies du discours de Sarkozy

Nous avons rappelé récemment, après d'autres, quelques approches différentes du traitement des co-occurrences (Mayaffre, 2008) : de l'extraction par le calcul des spécificités des co-occurrents d'un mot-pôle donné au repérage systématique de toutes les associations privilégiées du corpus, de la co-occurrence simple à la *poly-cooccurrence* (Martinez, 2003), la *Q-occurrence* (Massonie, 1986) ou la *cooccurrence généralisée* (Viprey, 1997), tout est aujourd'hui possible pour pressentir les réseaux thématiques et les isotopies d'un texte, aborder la textualité ou la texture, mettre à jour la cohérence ou le maillage d'un texte conçu comme un entrelacement lexical. Et nous renvoyons le lecteur à la thèse de Viprey (1997) pour la réflexion théorique la plus aboutie sur l'intérêt de la co-occurrence pour la linguistique textuelle.

Comme l'a montré récemment Brunet (2006, 2007 et 2008), le logiciel *Hyperbase* s'applique aujourd'hui à offrir plusieurs outils complémentaires pour embrasser le phénomène. Nous utiliserons, à propos de « patrie » chez Sarkozy, seulement l'outil le plus

récent implémenté dans le logiciel, la fonction « Associations » qui permet de constituer *des graphes de co-occurrents*.

L'idée de graphe n'est pas nouvelle puisque Demonet *et al* (1975) l'avaient préssentie, Heiden (2004) en fait un outil majeur du logiciel *Weblex* susceptible de produire un *lexicogramme* de l'ensemble des associations du corpus, et Véronis (2003 et 2004) propose, avec des applications instructives, l'outil *Hyperlex* à même de *cartographier* les co-occurrences du texte.

Dans cette lignée, et quoique de facture plus modeste, les graphes d'*Hyperbase* représentent une concrétion technique de l'idée de *réseaux lexicaux*. Pour cela, le logiciel met en forme le mot-pôle, ses principaux co-occurrents statistiques, mais encore les co-occurrents de ceux-ci (ou co-occurrents indirects), proposant ainsi une profondeur d'analyse à trois niveaux. Par là, le traitement statistique et sa mise en forme graphique donnent à voir des faisceaux isotopiques non triviaux qui se caractérisent, comme on le sait, par des phénomènes quantitatifs de récurrence[9] et d'échos sémantiques complexes que le lecteur peut percevoir dans la trame du texte (Figure 2).

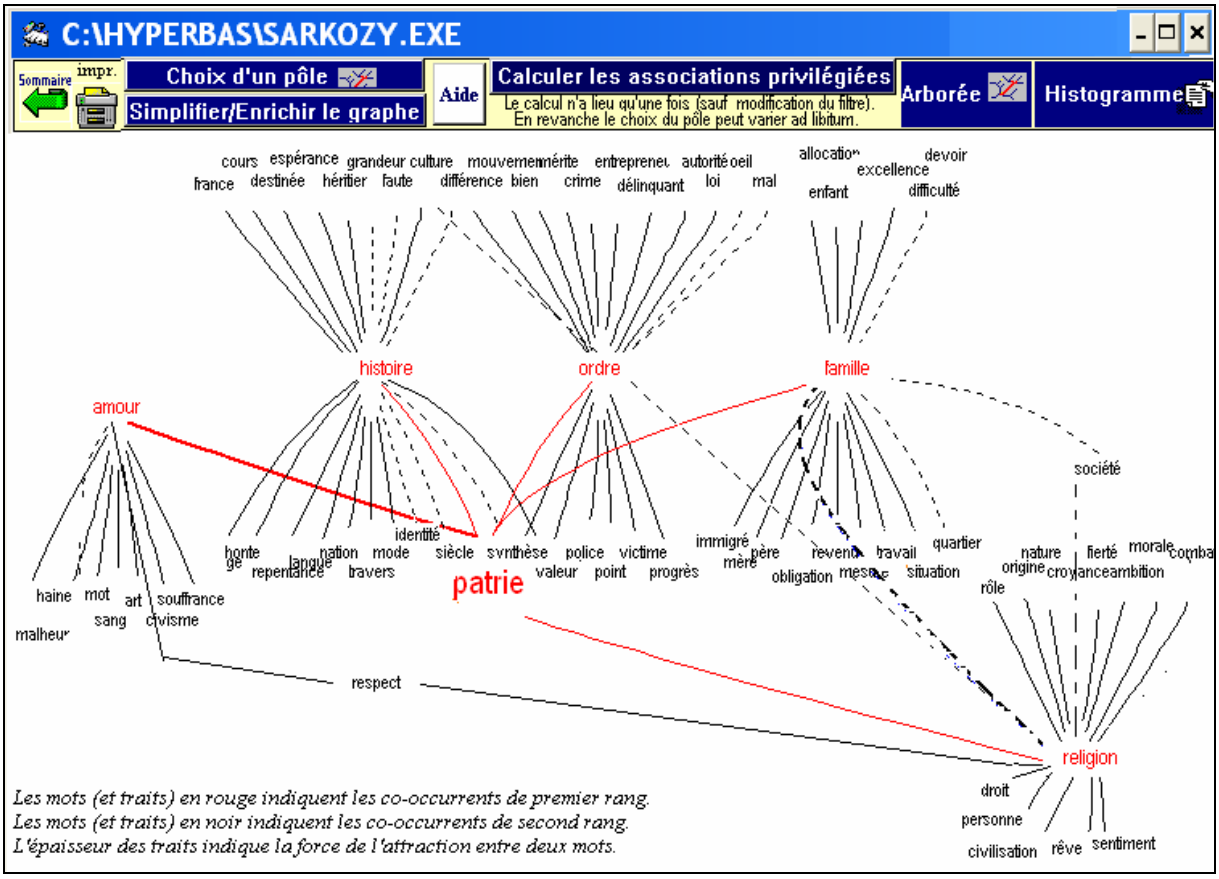

Figure 2 : Graphe des co-occurrents de « patrie » dans le discours de Sarkozy en 2007

[9] Dans les définitions qu'en donnent Greimas, Rastier, Arrivé ou Kerbrat, un phénomène isotopique est toujours produit par une *récurrence*, une *redondance*, des *reprises*, parfois des *itérations*. Qu'il nous soit donc permis de constater que l'isotopie fait partie de ces nombreux concepts qui impliquent, sans toujours se l'avouer, un traitement quantitatif. Dit plus directement : comment mesurer raisonnablement une *récurrence* significative sans lexicométrie ?

Une fois expurgé des mots outils et recentré sur les seuls substantifs, le traitement fait donc ressortir que « patrie » a 5 grands co-occurrents chez Sarkozy, qui marquent, si l'on veut bien considérer leurs co-occurrents respectifs, cinq dimensions du discours :

a) une dimension pathétique (« patrie » => « amour » –> « sang », « haine », « souffrance », etc.) comme dans cet exemple :

> Comment s'étonner qu'en dénigrant l'AMOUR de la PATRIE on réveille le nationalisme qui est la HAINE des autres ? Comment s'étonner que la mode exécrable de la repentance, en voulant faire expier aux Français les fautes supposées des générations passées, ressuscite des HAINES ancestrales que l'on croyait à tout jamais appartenir à l'histoire et rouvre des BLESSURES que le temps avait à peine commencé à fermer ? (Sarkozy, 18 mars 2007, meeting du Zénith à Paris).

De fait, le discours électoral de Sarkozy, au-delà du cas particulier de « patrie », est un discours qui joue autant sur l'émotion que sur la raison. Nous avons montré ailleurs (Mayaffre, 2007) qu'il se caractérise notamment par son extrémité lexicale (« haine », « détestation », « barbarie », « rêve », « excision » « voyou », etc.) là où le discours républicain classique est un discours de l'euphémisme lexical (« événement », « mouvement de protestation », « délinquant », etc.).

b) Une dimension historique/patriotique (« patrie » => « histoire » –> « France », « grandeur », « culture », « destinée » etc.) ; comme dans cet exemple :

> Je me fais une haute idée de la FRANCE, de ce qu'elle incarne aux yeux du monde, de son intelligence, de sa CULTURE, de sa vocation universelle. J'ai fait mienne son HISTOIRE. Pour moi il n'y a pas une HISTOIRE de France de gauche et une HISTOIRE de France de droite. Il n'y en a qu'une parce qu'il n'y a qu'une seule France. J'assume tout, je prends tout en partage et j'en suis fier. Je suis fier d'être un enfant de la PATRIE de Saint Louis, de Voltaire, de Victor Hugo, de Jaurès, de Blum, du Général de Gaulle, de Schuman, de Monnet. (Sarkozy, 11 février, meeting de Versailles).

c) Une dimension politique/autoritaire (« patrie => « ordre » –> « autorité », « délinquant », « crime », « police », etc.) ; comme ici :

> À bas l'AUTORITE ! Cela voulait dire : l'obéissance de l'enfant à ses parents, c'est fini ! Démodé ! La supériorité du maître sur l'élève, c'est fini ! Ringard ! La soumission à la LOI, c'est fini ! Dépassé ! Le pouvoir de POLICE, c'est fini ! Enfin ! Le respect de l'Etat et de ceux qui le représentent, c'est fini ! L'amour de la PATRIE, la fidélité à la France, à son drapeau, la gratitude vis-à-vis de ceux qui se sont battus pour elle, c'est fini ! La morale, c'est fini ! (Sarkozy, 23 février, meeting de Perpignan).

d) Une dimension sociétale/familiale (« patrie » => « famille » –> « père », « mère », mais aussi « travail », etc.) :

> Oui, à force de tout détester, la FAMILLE, la PATRIE, la religion, la société, le TRAVAIL, la politesse, la courtoisie, l'ordre, la morale. À force de tout détester, on finit par se détester soi-même. (Sarkozy, 5 avril 2007, meeting de Lyon).

Peu évoquée par les commentateurs, cette dimension familiale et éducative est l'un des aspects majeurs du discours de Sarkozy (que le calcul des spécificités par exemple révèle bien) et il est seulement étonnant de toucher à ce thème *via* les co-occurrences de « patrie ».

e) Enfin une dimension religieuse/spirituelle (« patrie » => « religion » –> « croyance », « rêve », etc.) que les discours du président Sarkozy sur Dieu et la foi le 20 décembre 2007 à Rome ou en janvier 2008 à Ryad révèleront plus encore.

La contextualisation de « patrie » par ses co-occurrents directs puis indirects permet donc sinon de donner un sens définitif au mot[10], en tout cas de retrouver des isotopies endogènes au corpus, assez loin des sentiers sur lesquels nous mènerait la définition dictionnairique. Par exemple, Sarkozy ne développe pas frontalement, en ce début du XXI^{e} siècle, la question de la patrie-territoire. « Patrie » est un élément isotopique complexe du discours, s'inscrivant dans plusieurs faisceaux distincts, le plus souvent mobilisé (et mobilisateur), dans le cadre de développements idéologiques généraux sur les valeurs comme l'ordre, l'autorité, la morale, le travail. Concrètement, concluons que statistiquement associé aux mots « amour » (co-occurrent direct) puis au mot « sang » (co-occurrent indirect) par exemple, ou encore associés au mot « famille », lui-même associé au mot « travail », « patrie » se contextualise ; ces associations co-occurrentielles constituent en elles-mêmes des contextes minimaux du corpus, ou des passages élémentaires, qui permettent de *lire* le texte, au sens plein, c'est-à-dire de produire du sens et de l'interprétation.

5 CONCLUSION

En définissant la co-occurrence comme forme minimale du contexte, cette contribution a essayé d'illustrer, de plusieurs points de vue, une idée unique : le passage d'une approche occurrentielle des corpus textuels à une approche co-occurrentielle ne représente pas seulement, pour la linguistique de corpus, un saut quantitatif mais une rupture qualitative.

Si la recherche d'occurrences renvoie à une démarche lexicographique ou encore à une linguistique logico-grammaticale dans laquelle il y aurait seulement des entités nucléaires indexées dans un dictionnaire et des règles de composition consultables dans une grammaire, la pratique des co-occurrences est d'essence contextualisante et ouvre d'autres perspectives. Elle réfléchit d'une part une science du vocabulaire *en usage* (l'usage minimal de *a* serait qu'il est, en corpus, statistiquement associé à *b*) et témoigne d'autre part de l'organisation textuelle en rendant compte, dans une perspective sémantique et herméneutique, de son maillage lexical.

En une formule proposée récemment par Valette (2008), la co-occurrence serait alors un élément essentiel d'une « *lexicologie textuelle* », c'est-à-dire d'une science du vocabulaire qui opérerait de manière endogène au texte pour contribuer à en rendre compte.

En cela, le traitement des co-occurrences, souvent abandonné au Traitement Automatique des Langues Naturelles (TALN) dans le cadre de la traductologie automatique ou de la désambiguïsation sémantique, est un enjeu majeur pour l'ADT et la lexicométrie qui, derrière ses techniques statistiques, espèrent être partie prenante des arts, des sciences et de l'interprétation des textes en proposant des parcours de lecture post-impressionistes à la fois originaux et contrôlés.

6 RÉFÉRENCES BIBLIOGRAPHIQUES

Adam, J.-M. (1999). *Linguistique textuelle. Des genres de discours aux textes*, Paris : Nathan.

Bejoin, H. & Thoiron Ph. (1992). « Macrostructure et microstructure dans un dictionnaire de collocations en langue de spécialité », *Terminologie et traduction*, vol. 2-3, p. 513-522.

Benson M., Benson E. & Ilson R. (1997). *The BBI Dictionary of English Word Combinations*, Amsterdam/Philadelphia : John Benjamins Publishing Company.

Brunet E. (2006). « Navigation dans les rafales », *in :* Viprey J.-M. (éd.), *JADT'06*, Besançon : Presses Universitaires de Franche-Comté, vol. I, p. 15-29.

[10] On a compris au terme de cet article, que, précisément, nous ne croyons ni à la singularité du sens ni à une acception définitive des mots. Les mots sont tous potentiellement polysémiques (notamment en diachronie) et l'étude de leur usage témoigne de cette pluralité sémantique.

Brunet E. (2007). « Séquences et fréquences. Mises en œuvre dans Hyperbase », *Lexicométrica, Topographie et topologie textuelles* [http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9/brunet.pdf]
Brunet E. (2008). « Les séquences (suite) », *in* : Heiden S. (éd.), *Actes des JADT 2008* (sous presse).
Church K. W. & Hanks P. (1990). « Word Association Norms, Mutual Information, And Lexicography », *Computational Linguistics*, vol. 16(1), p. 177-210.
Daille B., Williams G. (dir.) (2001). *Collocation : computational extraction, analysis and exploitation*, Workshop at the 39th Annual Meeting and 10th Conference of the European Chapter of ACL.
Demonet M. *et al.* (1975). *Des tracts en mai 68*, Paris : Colin.
Dugast D. (1979). V*ocabulaire et discours : fragments de lexicologie quantitative : essai de lexicométrie organisationnelle*, Genève-Paris : Slatkine-Champion.
Dunning, T. (1993). « Accurate Methods for the Statistics of surprise and Coincidence », *Computational Linguistics*, vol. 19(1), p. 61-74.
Eluerd R. (2000). *La lexicologie*, Paris : PUF.
Fung P. & McKeown K. (1997). « A technical word and term translation aid using noisy parallel corpora across language groups », *Machine Translation*, vol. 12(1-2), p. 53-87.
Firth J. (1957). « A Synopsis of Linguistic Theory 1930-1955 », *Studies in Linguistic Analysis*, p. 1-32.
Gross G. (1996). *Les expressions figées en français : noms composés et autres locutions*, Paris : Ophrys.
Guilhaumou J. & Monnier R. (textes réunis par…) (2006). *Dictionnaire des usages socio-politiques (1770-1815)*, Fascicule 8 : « Patrie, patriotisme. Notions pratiques », Paris : Honoré Champion.
Guiraud P. (1960). *Problèmes et méthodes de la statistique linguistique*, Paris : Larousse.
Haussmann F. J. (1979). « Un dictionnaire des collocations est-il possible ? », *Travaux de linguistique et de littérature*, vol. XVII(l), p. 187-195.
Heiden S. (2004). « Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex », in : Purnelle G. (ed.), *JADT 2004-Le poids des mots*, Louvain : Presses universitaires de Louvain, p. 577-588.
Harris Z. S. (1957). « Co-occurrence and transformation in linguistic structure », *Language*, n°33, p. 283-340.
Heiden S. et Lafon P. (1998). « Cooccurrences. La CFDT de 1973 à 1992 », in : *Des mots en liberté, Mélanges Maurice Tournier*, Paris : ENS Éditions, tome 1, p. 65-83.
Ilgenfritz, P. *et al.* (1989). *Langenscheidts Kontextwörterbuch Französisch-Deutsch. Ein neues Wörterbuch zum Schreiben, Lernen, Formulieren*, Berlin-München : Langenscheidt.
Lafon P. (1984). *Dépouillements et Statistiques en Lexicométrie*, Genève-Paris : Slatkine-Champion.
Laks B. (2008). « Pour une phonologie de corpus », *Journal of French Language Studies* (sous presse).
Lebart L. & Salem, A. (1994), *Statistique textuelle*, Paris : Dunod.
Longrée D., Luong X. & Mellet S. (2008). « Les motifs : un outil pour la caractérisation topologique des textes », in : Heiden S. (éd.), *Actes des JADT 2008* (sous presse).
Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, Cambridge (Massachusetts) : The MIT Press.
Martinez W. (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse de Doctorat, Université de la Sorbonne nouvelle-Paris 3, sous la direction d'A. Salem.
Massonie J.-P. (1986). *Pratique de l'analyse des correspondances*, Besançon : Annales Littéraires de l'Université de Franche-Comté.
Mayaffre D. (2002a). « L'Herméneutique numérique », *L'Astrolabe. Recherche littéraire et Informatique*, (revue électronique).
Mayaffre D. (2002b). « Les corpus réflexifs : entre architextualité et hypertextualité », *Corpus*, 1, p. 51-69.
Mayaffre D. (2006). « Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques », in : Rastier F. et Ballabriga M. (éds), *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation*, Toulouse PUT, p. 15-26. (Lire en ligne sur *Texto ! Textes et cultures*).

Mayaffre D. (2007). « Vocabulaire et discours électoral de Sarkozy : entre modernité et pétainisme », *La Pensée*, 352, p. 65-80.
Mayaffre D. (2008). « Quand 'travail', 'famille', 'patrie' co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence », in : Heiden S. (éd.), *Actes des JADT 2008* (sous presse).
Mel'Cuk, I. *et al.* (1984). *Dictionnaire Explicatif et Combinatoire du français contemporain I*, Montréal : Presses de l'Université de Montréal.
Mellet S. et Barthélemy J.-P. (2007). « La topologie textuelle : légitimation d'une notion émergente », *Lexicométrica*, numéro thématique « Topographie et topologie textuelle » (http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9/mellet.pdf.)
Rastier F. (2001). *Arts et sciences du texte*, Paris : PUF.
Rastier F. (2007). « Passages », *Corpus*, 6, p. 25-54.
Rastier F. et Valette M. (*à paraître*). « De la polysémie à la néosémie », *Langue française*.
Salem A. (1993). *Méthodes de la statistique textuelle*, Thèse d'Etat, Paris 3.
Saussure F. de (éd 1995), *Cours de linguistique générale*, Paris : Payot.
Tournier M. (1980). « D'où viennent les fréquences de vocabulaire ? La lexicométrie et ses modèles », *Mots*, 1, p. 189-209.
Valette M. (2008). « À quoi servent les lexiques sémantiques généralistes ? Discussion et propositions », *Cahiers du Cental* (sous presse).
Véronis J. (2003). « Cartographie lexicale pour la recherche d'information », in : *Actes de TALN 2003*, p. 265-274.
Véronis, J. (2004). « Hyperlex : lexical cartography for information retrieval », *Computer, Speech and Language*, vol. 18/3, p. 223-252.
Viprey J.-M. (1997). *Dynamique du vocabulaire des Fleurs du mal*, Paris Honoré Champion.
Viprey, J.-M. (2005a). « Philologie numérique et herméneutique intégrative », in : Adam J.-M. & Heidmann U. (eds.), *Sciences du texte et analyse de discours*, Genève : Slatkine, p. 51-68.
Viprey J.-M. (2005b). « Corpus et sémantique discursive : éléments de méthode pour la lecture des corpus », in : A. Condamines (dir.), *Sémantique et corpus*, Paris : Lavoisier, p. 245-276.
Viprey J.-M. (2006). « Structure non-séquentielle des textes », *Langages*, n°163, p. 71-85.
Williams G. (1999). *Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique*, Thèse de doctorat, Université de Nantes.
Wu H. & Zhou M. (2003). « Synonymous collocation extraction using translation information », in : Hinrichs E. & Roth D. (eds.), *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo : National Institue of Informatics, p. 120-127.